Emma Pierson
@2plus2make5

Check out our new paper to be presented at FaaCT, "Trucks Don't Mean Trump: Diagnosing Human Error in Image Analysis" led by @jdzamfi with coauthors Jerry Chen, Emily Wen, @allisonkoe, @NikhGarg, and me (arxiv.org/abs/2205.07333)! 1/

arxiv.org
Trucks Don't Mean Trump: Diagnosing Human Error in Ima...
Algorithms provide powerful tools for detecting and dissecting human bias and error. Here, we develop machin...

2:16 PM · May 17, 2022

2 12 32 6

Reply

Emma Pierson @2plus2make5 · May 17, 2022
People often have to analyze images in high-stakes settings - medicine, content moderation, etc. Here, we develop a machine learning approach to analyze the ways in which they err in doing so. 2/

1 3

Emma Pierson @2plus2make5 · May 17, 2022
We rely on a unique dataset kindly provided to us by the @nytimes: 16 million human predictions of whether a neighborhood voted for Trump or Biden in the 2020 election, based on a Google Street View image (nytimes.com/interactive/20...). 3/

1 2

Emma Pierson @2plus2make5 · May 17, 2022
This data is cool because it has a large number of human judgments on each image (more than a thousand!) and a ground truth defined independent of human judgment. The latter is unusual, and key - otherwise studying human error would be circular. 4/

1 3

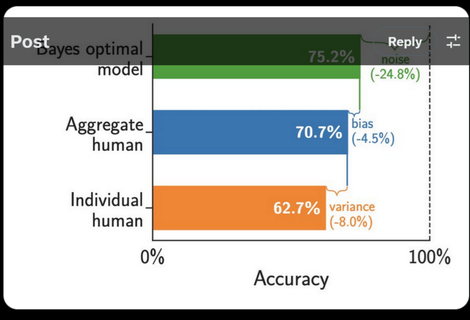
Emma Pierson @2plus2make5 · May 17, 2022
We show that by training a machine learning estimator of p(voted Trump | neighborhood image), you can get a number of useful results... 5/

1 2

Emma Pierson @2plus2make5 · May 17, 2022
You can decompose human error into bias, variance, and noise terms, analogous to the decomposition for ML classifiers. Bias = suboptimality in the aggregate human decision; variance = suboptimality due to variance across individual humans; noise = unavoidable error. 6/

1 1

Emma Pierson @2plus2make5 · May 17, 2022
In our data, accuracy loss due to bias is smaller than that due to variance + noise - humans in aggregate are actually pretty good at this task (wisdom of crowds) even though individual humans are erratic and the task is hard. 7/



1 3

Emma Pierson @2plus2make5 · May 17, 2022
This decomposition is actionable. If, eg, doctors mainly lose accuracy due to bias, we might retrain them; if they are accurate in aggregate but individually high-variance, we may need second opinions; and if the images are noisy, we may need an alternate diagnostic modality. 8/

1 3

Emma Pierson @2plus2make5 · May 17, 2022
We also provide several methods for identifying specific features which lead people astray, like pickup trucks (people think they indicate Trump more than they really do). 9/

1 1

Emma Pierson @2plus2make5 · May 17, 2022
Even if our estimate of p(voted Trump | neighborhood image) is imperfect, we show our approach can still provide useful insights into human error as long as our machine learning model adds signal beyond human judgment, a property we verify. 10/

1 2

Emma Pierson @2plus2make5 · May 17, 2022
You can follow our playbook on other image (or non-image) datasets with human (or algorithmic) judgments + objectively defined ground truth - and those datasets are becoming increasingly available! See eg docs.nightingalescience.org for medical datasets with objective ground truth.

2